

Brand	Price (\$)	Score
Bose	180	76
Skullcandy	150	71
Koss	95	61
Phillips/O'Neill	70	56
Denon	70	40
JVC	35	26

- Compute SST, SSR, and SSE.
- Compute the coefficient of determination r^2 . Comment on the goodness of fit.
- What is the value of the sample correlation coefficient?

19. In exercise 7 a sales manager collected the following data on x = annual sales and y = years of experience. The estimated regression equation for these data is $\hat{y} = 80 + 4x$.

DATA file
Sales

Salesperson	Years of Experience	Annual Sales (\$1000s)
1	1	80
2	3	97
3	4	92
4	4	102
5	6	103
6	8	111
7	10	119
8	10	123
9	11	117
10	13	136

- Compute SST, SSR, and SSE.
 - Compute the coefficient of determination r^2 . Comment on the goodness of fit.
 - What is the value of the sample correlation coefficient?
20. *Bicycling*, the world's leading cycling magazine, reviews hundreds of bicycles throughout the year. Their "Road-Race" category contains reviews of bikes used by riders primarily interested in racing. One of the most important factors in selecting a bike for racing is the weight of the bike. The following data show the weight (pounds) and price (\$) for 10 racing bikes reviewed by the magazine (*Bicycling* website, March 8, 2012).

DATA file
RacingBicycles

Brand	Weight	Price (\$)
FELT F5	17.8	2100
PINARELLO Paris	16.1	6250
ORBEA Orca GDR	14.9	8370
EDDY MERCKX EMX-7	15.9	6200
BH RC1 Ultegra	17.2	4000
BH Ultralight 386	13.1	8600
CERVELO S5 Team	16.2	6000
GIANT TCR Advanced 2	17.1	2580
WILIER TRIESTINA Gran Turismo	17.6	3400
SPECIALIZED S-Works Amira SL4	14.1	8000

- Use the data to develop an estimated regression equation that could be used to estimate the price for a bike given the weight.

DATA *file*
GoldHoldings

Country	Gold Value (\$ billions)	Debt (% of GDP)
China	63	17.7
France	146	81.7
Germany	203	83.2
Indonesia	33	69.2
Italy	147	119.0
Netherlands	36	63.7
Russia	50	9.9
Switzerland	62	55.0
United States	487	93.2

- Develop a scatter diagram for the total value of a country's gold holdings (\$ billions) as the independent variable.
- What does the scatter diagram developed in part (a) indicate about the relationship between the two variables? Does there appear to be any outliers and/or influential observations? Explain.
- Using the entire data set, develop the estimated regression equation that can be used to predict the debt of a country given the total value of its gold holdings.
- Suppose that after looking at the scatter diagram in part (a) that you were able to visually identify what appears to be an influential observation. Drop this observation from the data set and fit an estimated regression equation to the remaining data. Compare the estimated slope for the new estimated regression equation to the estimated slope obtained in part (c). Does this approach confirm the conclusion you reached in part (d)? Explain.

54. The following data show the annual revenue (\$ millions) and the estimated team value (\$ millions) for the 30 Major League Baseball teams (*Forbes* website, January 16, 2014).

DATA *file*
MLBValues

Team	Revenue (\$ millions)	Value (\$ millions)
Arizona Diamondbacks	195	584
Atlanta Braves	225	629
Baltimore Orioles	206	618
Boston Red Sox	336	1312
Chicago Cubs	274	1000
Chicago White Sox	216	692
Cincinnati Reds	202	546
Cleveland Indians	186	559
Colorado Rockies	199	537
Detroit Tigers	238	643
Houston Astros	196	626
Kansas City Royals	169	457
Los Angeles Angels of Anaheim	239	718
Los Angeles Dodgers	245	1615
Miami Marlins	195	520
Milwaukee Brewers	201	562
Minnesota Twins	214	578
New York Mets	232	811
New York Yankees	471	2300
Oakland Athletics	173	468
Philadelphia Phillies	279	893

54

Team	Revenue (\$ millions)	Value (\$ millions)
Pittsburgh Pirates	178	479
San Diego Padres	189	600
San Francisco Giants	262	786
Seattle Mariners	215	644
St. Louis Cardinals	239	716
Tampa Bay Rays	167	451
Texas Rangers	239	764
Toronto Blue Jays	203	568
Washington Nationals	225	631

- Develop a scatter diagram with Revenue on the horizontal axis and Value on the vertical axis. Looking at the scatter diagram, does it appear that there are any outliers and/or influential observations in the data?
- Develop the estimated regression equation that can be used to predict team value given the value of annual revenue.
- Use residual analysis to determine whether any outliers and/or influential observations are present. Briefly summarize your findings and conclusions.

14.10 Practical Advice: Big Data and Hypothesis Testing in Simple Linear Regression

In Chapter 7, we observed that the standard errors of the sampling distributions of the sample mean $\bar{x}$ (shown in formula 7.2) and the sample proportion of $\bar{p}$ (shown in formula 7.5) decrease as the sample size increases. In Chapters 8 and 9, we observed that this results in narrower confidence interval estimates for μ and p and smaller p -values for the tests of the hypotheses $H_0: \mu \leq \mu_0$ and $H_0: p \leq p_0$ as the sample size increases. These results extend to simple linear regression. In simple linear regression, as the sample size increases,

- the p -value for the t -test used to determine whether a significant relationship exists between the dependent variable and the independent decreases;
- the confidence interval for the slope parameter associated with the independent variable narrows;
- the confidence interval for the mean value of y narrows;
- the prediction interval for an individual value of y narrows.

Thus, we are more likely to reject the hypothesis that a relationship does not exist between the dependent variable and the independent variable and conclude that a relationship exists as the sample size increases. The interval estimates for the slope parameter associated with the independent variable, the mean value of y , and predicted individual value of y will become more precise as the sample size increases. But this does not necessarily mean that these results become more reliable as the sample size increases.

No matter how large the sample used to estimate the simple linear regression equation, we must be concerned about the potential presence of nonsampling error in the data. It is important to carefully consider whether a random sample of the population of interest has actually been taken. If the data to be used for testing the hypothesis of no relationship between the independent and dependent variable are corrupted by nonsampling error, the likelihood of making a Type I or Type II error may be higher than if the sample data are free